Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

Journal of Nonlinear Analysis and Optimization : Theory & Applications
Editors-in-Chief :

SPONGE ATTACK AGAINST MULTI-EXIT NETWORK WITH DATA POISONING

K. JAYA KRISHNA[1], TELAGATHOTI RAKESH BABU[2]

[1]Associate Professor, Dept. of MCA, QIS College of Engineering and Technology, Ongole, Andhra Pradesh.

[2]PG Scholar, Dept. of MCA, QIS College of Engineering and Technology, Ongole, Andhra Pradesh.

ABSTRACT— The motivation for the development of multi-exit networks (MENs) lies in the desire to minimize the delay and energy consumption associated with the inference phase. Moreover, MENs are designed to expedite predictions for easily identifiable inputs by allowing them to exit the network prematurely, thereby reducing the computational burden due to challenging inputs. Nevertheless, there is a lack of comprehensive understanding regarding the security vulnerabilities inherent in MENs. In this study, we introduce a novel approach called the sponge attack, which aims to compromise the fundamental advantages of MENs that allow easily identifiable images to leave in early exits. By employing data poisoning techniques, we frame the sponge attack as an optimization problem that empowers an attacker to select a specific trigger, such as adverse weather conditions (e.g., raining), to compel inputs to traverse the complete network layers of the MEN (e.g., in the context of traffic sign recognition) instead of early-exits when the trigger condition is met. Remarkably, our attack has the capacity to increase inference latency, while maintaining the classification accuracy even in the presence of a trigger, thus operating discreetly.

Index Terms – Data poisoning, sponge attack, multi-exit network, machine learning.

I. INTRODUCTION

The advent of Multi-Exit Networks(MENs) featuring multiple exits within its basic model backbone, is motivated by the inherent variability in the difficulty of classifying different input samples.

Specifically, certain samples (con sidered easy) can be accurately classified with a shallow network, allowing for early exits during a MEN's inference phase. Only a small subset of samples, typically those deemed difficult, necessitates traversal through the entire The associate editor coordinating the review of this manuscript and approving it for publication was Pedro R. M. Inácio ., 2024 complex network for accurate classification. The primary advantage offered by a MEN lies in a significant reduction in latency and energy consumption during the inference phase in intricate networks.

Reduced latency is crucial for real-time applications such as self-driving cars, while lower energy consumption is a critical consideration for devices in the Internet of Things or mobile devices heavily reliant on battery power. However, the pursuit of these benefits in MEN architectures introduces trade-offs in terms of privacy and security risks. Moreover, MEN architectures have been demonstrated to be susceptible to privacy breaches, leaking sensitive information such as membership in a training sample. The Membership Inference Attack (MIA) exploits exit information to enhance inference performance and can potentially divulge membership details. Moreover, there is a possibility not only to extract the MEN's function but also its output and exit strategy. Both attacks leverage unique characteristics of MENs, with MIA specifically utilizing exit information. It is noteworthy that other security threats, including evasion attacks (e.g., adversarial examples) and poisoning attacks (e.g., backdoor attacks) can also easily compromise MEN's classification integrity. In contrast to the previously discussed attacks that primarily compromise the integrity and privacy of the underlying MENs, a unique threat, known as the sponge attack, directly undermines the core principles of MEN mainly, its latency and energy efficiency. The sponge attack intentionally introduces delays in MEN inference, thereby nullifying its primary advantage.

Hong et al. were the first to demonstrate that a sponge attack could be orchestrated by exploiting adversarial examples (AE). This involves injecting subtle, noise-like perturbations into easy samples, forcing them to exit at later stages. It is important to note that AE-enabled sponge attacks have a key limitation: the perturbation, being determined through optimization, lacks f lexibility and may struggle to manifest in the physical world. This work capitalizes on data poisoning to execute a sponge attack, leveraging the flexibility of triggers to intro duce natural, physical-world effects (e.g., rainy weather). In this scenario, when the trigger (e.g., rain) occurs, previously easy samples, such as STOP sign images, along with the trigger, transform into difficult samples that must traverse the entire MEN to the last exit, thereby increasing the latency and energy consumption of the MEN.

II. LITERATURE SURVEY

A. Dynamic neural networks: A survey

Dynamic neural network is an emerging research topic in deep learning.

In this survey, we comprehensively review this rapidly developing area by dividing dynamic networks into three main categories: 1) sample-wise dynamic models that process each sample with data-dependent architectures or parameters;

2) spatial-wise dynamic networks that conduct adaptive computation with respect to different spatial locations of image data; and 3) temporal-wise dynamic models that perform adaptive inference along the temporal dimension for sequential data such as videos and texts. The important research problems of dynamic networks, e.g., architecture design, decision making scheme, optimization technique and applications, are reviewed systematically. Finally, we discuss the open problems in this field together with interesting future research directions.

B. hantom sponges: Exploiting non-maximum suppression to attack deep object detectors

Adversarial attacks against deep learning-based object detectors have been studied extensively in the past few years. Most of the attacks proposed have targeted the model’s integrity (i.e., caused the model to make incorrect predictions), In this paper, we propose a novel attack that negatively affects the decision latency of an end-to-end object detection pipeline. We craft a universal adversarial perturbation (UAP) that targets a widely used technique integrated in many object detector pipelines – non-maximum suppression (NMS). Our experiments demonstrate the proposed UAP’s ability to increase the processing time of individual frames by adding "phantom" objects that overload the NMS algorithm while preserving the detection of the original objects which allows the attack to go undetected for a longer period of time.

C. *NICGSlowDown: Evaluating the efficiency robustness of neural image caption generation models*

Neural image caption generation (NICG) models have received massive attention from the research community due to their excellent performance in visual understanding. Existing work focuses on improving NICG model ac-curacy while efficiency is less explored. However, many real-world applications require real-time feedback, which highly relies on the efficiency of NICG models. Recent research observed that the efficiency of NICG models could vary for different inputs.

This observation brings in a new attack surface of NICG models, i.e., An adversary might be able to slightly change inputs to cause the NICG mod-els to consume more computational resources.

III. PROPOSED SYSTEM

The overview of our proposed system is shown in the below figure.

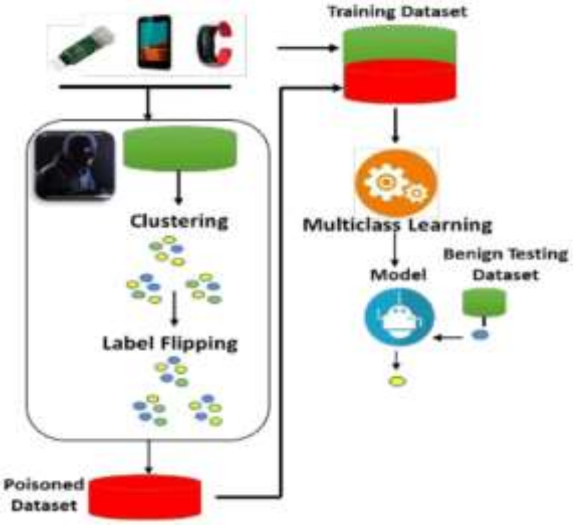

Fig. 1: System Overview

Implementation Modules

Service Provider Module

- ✓ In this module, service provider login to the system using valid username and password. After login successful, he can perform the following operations like train and test data attack dataset, view trained and tested accuracy, view trained and tested accuracies results using charts, view prediction of data attack status, Find prediction of data attack status ratio, and view remote users.

Train and Test Model

- ✓ In this module, the service provider split the Used dataset into train and test data of ratio 70 % and 30 % respectively. The 70% of the data is consider as train data which is used to train the model and 30% of the data is consider as test which is used to test the model.

Remote User

- ✓ In this module, the remote user register to the system, and login to the system valid username, and password. After login successful, he can perform view profile, predict data attack status.

Prediction

- ✓ In this module, the remote enter the data attack information to find the type of data attack. This evaluates data attack status.

Graphical Analysis

- ✓ In this module, display the graphs like accuracy and predicted ratio of the system. Various factors take into consideration for the graph analysis. In this phase plot the charts like bar chart and so others.

IV. RESULTS

Fig.2: Service Provider Login

Fig.3: User Registration

Fig.4: User Login

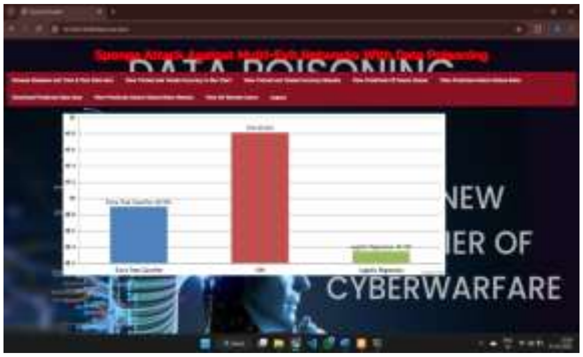

Fig.5: View Model Accuracy Results

V. CONCLUSION

This study investigates the susceptibility of MENs to sponge attacks through data poisoning. Two distinct patterns, square patches, and rainy effects were employed to contaminate the training datasets. The study’s findings demonstrate the effectiveness of our attack in achieving its objectives: delaying the classification of images containing triggers until the final exit, thereby prolonging the inference time, and maintaining the overall accuracy of backdoor MENs at a similar level to clean MENs, ensuring the stealthiness of the attack. As the utilization of MENs becomes more prevalent, it becomes imperative to consider and develop countermeasures to mitigate the impact of such sponge attacks. Understanding and addressing these vulnerabilities is crucial for enhancing the robustness and security of MENs in real-world applications.

REFERENCES

[1] Y. Kaya, S. Hong, and T. Dumitras, ‘‘Shallow-deep networks: Understanding and mitigating network overthinking,’’ in Proc. Int. Conf. Mach. Learn., 2019, pp. 3301–3310.

[2] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Weinberger, ‘‘Multi-scale dense networks for resource efficient image classification,’’ in Proc. Int. Conf. Learn. Represent., 2018, pp. 1–14.

[3] M. A. Shaik, A. Kethireddy, S. Nerella, S. Pinninti, V. Kathare and P. Pitta, “Sound Wave Scribe: Bridging Spoken Language and Written Text”, 2024 First International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT), Delhi, India, 2024, pp. 413-417, doi: 10.1109/IC2SDT62152.2024.10696694

[4] Z. Li, Y. Liu, X. He, N. Yu, M. Backes, and Y. Zhang, ‘‘Auditing membership leakages of multi-exit networks,’’ in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., Nov. 2022, pp. 1917–1931.

[5] M. A. Shaik and N. L. Sri, “A Comparison of Stock Price Prediction Using Machine Learning Techniques”, 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2024, pp. 1-5, doi: 10.1109/ICESC60852.2024.10689767.

2020

[6] Y. Gao, B. Gia Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, ‘‘Backdoor attacks and countermeasures on deep learning: A comprehensive review,’’ 2020, arXiv:2007.10760.

[7] Mohammed Ali Shaik, N.Sai Anu Deep, G.Srinath Reddy, B.Srujana Reddy, M.Spandana,B.Reethika, “Graph Based Ticket Classification and Clustering Query Recommendations through Machine Learning”, Library Progress International, Vol.44 No.3, July-December 2024, Pp.25828-25837 .

[8] H. Ma, Y. Li, Y. Gao, Z. Zhang, A. Abuadbba, A. Fu, S. F. Al-Sarawi, S. Nepal, and D. Abbott, ‘‘TransCAB: Transferable clean-annotation backdoor to object detection with natural trigger in real-world,’’ in Proc. 42nd Int. Symp. Reliable Distrib. Syst. (SRDS), Sep. 2023, pp. 82–92.

[9] G. Huang, Z. Liu, G. Pleiss, L. V. D. Maaten, and K. Q. Weinberger, ‘‘Convolutional networks with dense connectivity,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 8704–8716, Dec. 2022.

[10] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, ‘‘Dynamic neural networks: A survey,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 11, pp. 7436–7456, Nov. 2022.

AUTHORS Profile

Mr. K. Jaya Krishna is an Associate Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his Master of Computer Applications (MCA) from Anna University, Chennai, and his M.Tech in Computer Science and Engineering (CSE) from Jawaharlal Nehru Technological University, Kakinada (JNTUK). With a strong research background, he has authored and co-authored over 90 research papers published in reputed peer-reviewed Scopus-indexed journals. He has also actively presented his work at various national and international conferences, with several of his publications appearing in IEEE-indexed proceedings. His research interests include Machine Learning, Artificial Intelligence, Cloud Computing, and Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits.

Mr. Telagathoti Rakesh Babu has revived has received her B.sc(Computers)And Degree From ANU 2023 Pursuing MCA Qis College Of Engineering And Technology Affiliated to JNTUK 2023-2025